LETTER

doi:10.1038/nature10995

Evidence of non-random mutation rates suggests an evolutionary risk management strategy

Iñigo Martincorena[1], Aswin S. N. Seshasayee[1]† & Nicholas M. Luscombe[1,2,3,4]

A central tenet in evolutionary theory is that mutations occur randomly with respect to their value to an organism; selection then governs whether they are fixed in a population. This principle has been challenged by long-standing theoretical models predicting that selection could modulate the rate of mutation itself[1,2]. However, our understanding of how the mutation rate varies between different sites within a genome has been hindered by technical difficulties in measuring it. Here we present a study that overcomes previous limitations by combining phylogenetic and population genetic techniques. Upon comparing 34 *Escherichia coli* genomes, we observe that the neutral mutation rate varies by more than an order of magnitude across 2,659 genes, with mutational hot and cold spots spanning several kilobases. Importantly, the variation is not random: we detect a lower rate in highly expressed genes and in those undergoing stronger purifying selection. Our observations suggest that the mutation rate has been evolutionarily optimized to reduce the risk of deleterious mutations. Current knowledge of factors influencing the mutation rate—including transcription-coupled repair and context-dependent mutagenesis—do not explain these observations, indicating that additional mechanisms must be involved. The findings have important implications for our understanding of evolution and the control of mutations.

The question of whether spontaneous mutations occur randomly has attracted great interest for decades[1–6]. The answer has fundamental implications for evolutionary theory as well as for our understanding of pathogen evolution and certain human diseases. Mutation rates can be modulated by at least three selective forces that can vary in strength depending on genomic location: the cost of deleterious mutations, the need for adaptive mutations, and the cost of fidelity in DNA replication and repair[4,6]. Thus, although mutations are generally assumed to occur independently of their fitness effect, it is conceivable that local mutation rates themselves might evolve, resulting in genomes whose mutations occur non-randomly: more frequently where they are more often advantageous and less frequently where they are most deleterious. Currently, however, there is little evidence that local mutation rates have been optimized during evolution, with the limited exceptions of bacterial contingency loci and somatic hypermutation in the vertebrate immune system.

Our understanding of how the mutation rate varies along a genome has been restricted by the lack of reliable approaches to measure local mutation rates on a large scale. Experimentally, absolute mutation rates can be determined using gene reporters in fluctuation tests, but these are unsuitable for measurements in native genes. Alternatively, in theory, relative mutation rates can be estimated from the accumulation of mutations at selectively neutral positions. Indeed, synonymous or non-coding sites are often used as proxies[7,8] and intriguing correlations between local synonymous substitution rates and gene function or fitness cost have been reported[8,9]. However, because selection can act on these sites through factors like codon-usage preference, RNA-folding stability and *cis*-regulatory elements, interpretation of these observations in support of optimized mutation rates has remained contentious.

Traditional studies comparing two species suffer from the fundamental limitation that the effects of selection and mutation rate on sequence divergence cannot be distinguished. However, as the two processes leave distinct patterns of polymorphisms, population genetic techniques can be applied to disentangle their relative contributions[10–12]. In examining more than 120,000 single-nucleotide polymorphisms across 34 *E. coli* strains, we exploited this fact to quantify the heterogeneity of the local mutation rate and to test for evidence of evolutionary optimization. Multiple alignments of 2,930 orthologous genes passed stringent phylogenetic filters checking for artefacts like interspecies gene transfer, orthologue misidentification, sequencing errors and misalignments (Supplementary Information, section 2.2). Using these alignments, we then calculated the synonymous diversity of each gene ($\theta_s = 2N\mu$) (Methods and Supplementary Information, section 2.1).

Figure 1a immediately highlights the marked variability in the density of synonymous polymorphisms along the *E. coli* genome. Although the mean synonymous diversity (θ_s) is 0.04, these values vary by more than an order of magnitude from less than 0.002 to more than 0.10 depending on the gene (Fig. 1a–c, f). The observed heterogeneity is much larger than expected under a regime of uniform mutation rate alone (Fig. 1f and Supplementary Information, section 6.1).

The synonymous diversity also shows signs of regional organization (Fig. 1b, c). Mutational hot and cold regions often span entire operons, indicating that gene function has a large effect on θ_s. Neighbouring genes tend to have similar θ_s values, particularly when they are encoded on the same strand and share common functions (Fig. 1g). A weaker correlation is also apparent for genes with distinct functions and on opposing strands, suggesting that synonymous diversity is affected by regional factors at a resolution of few kilobases.

Although the observed variation in θ_s is suggestive, selective and non-selective processes other than mutation rate can also influence synonymous diversity. Synonymous sites in *E. coli* are known to experience purifying selection. The best evidence for this is the negative correlation between inter-species synonymous divergence (dS) and codon usage bias (CUB) ($R^2 = 0.31$ for *E. coli*–*Salmonella enterica*, Supplementary Fig. 7)[13]. In contrast, within-species diversity in *E. coli* (θ_s) does not correlate with CUB ($R^2 < 0.01$, Fig. 2a) demonstrating that, contrary to dS, the variation in θ_s is largely neutral with respect to codon usage.

To explain this unexpected result, we applied a mutation-selection-drift model to estimate the strength of selection acting on codon usage[14] (Supplementary Information, section 3.1.1.2). In agreement with previous reports[15], we find that this is relatively weak for most genes ($\gamma_{mean} = -0.41$, $\gamma_{std} = 0.40$, Fig. 2b). These estimated selection coefficients are too small to affect within-species θ_s substantially, but large enough to affect inter-species dS[14] (Fig. 2a and Supplementary Fig. 8). The other prominent source of selection at synonymous sites,

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK. [2]Okinawa Institute of Science & Technology, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan. [3]UCL Genetics Institute, Department of Genetics, Environment and Evolution, University College London, Gower Street, London WC1E 6BT, UK. [4]Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK. †Present address: National Centre for Biological Sciences, TIFR, GKVK, Bellary Road, Bangalore 560065, India.

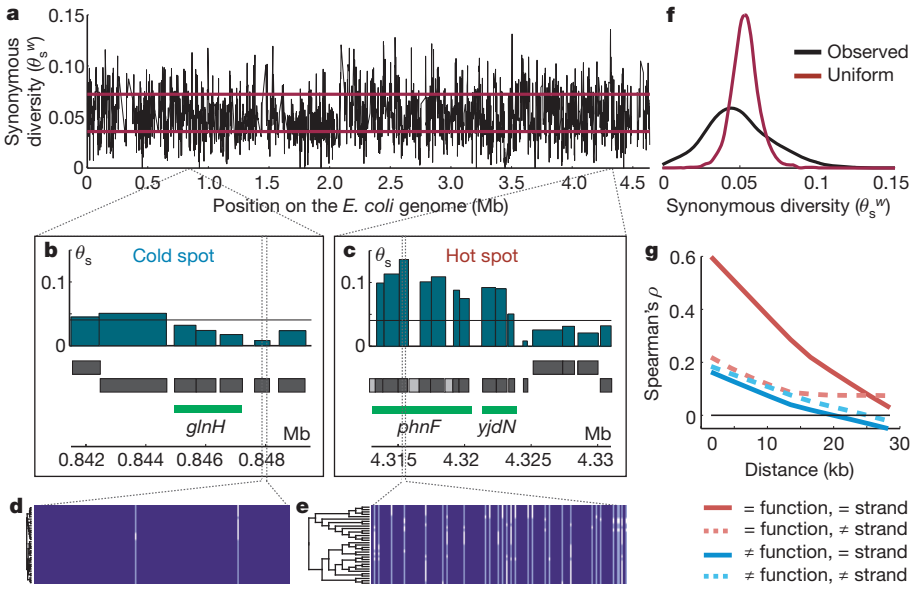

Figure 1 | Synonymous diversity along the *E. coli* genome is heterogeneous. **a**, Distribution of the 2,930 θ_s values (using θ_s^W, see Supplementary Information, section 6.1) plotted along the *E. coli* K12 MG1655 genome showing that synonymous diversity is highly variable. The red lines indicate the upper and lower 2.5% limits of the expected distribution under a regime of uniform mutation rate. Forty-one per cent of θ_s^W values fall outside of the expected range. **b**, **c**, Details of two mutational hot and cold regions displaying the genomic coordinates, annotated genes and their respective θ_s values. Each gene is shown as a black box on the sense or anti-sense strands (grey for genes with no θ_s available). Green horizontal lines indicate operons, labelled by the name of their lead gene. **d**, **e**, Multiple-sequence alignments of 180-base-pair segments of two genes, (**d**) *dps* ($\theta_s = 0.008$; $TD_{syn} = 0.35$; RNA-Seq level = 139.9), and (**e**) *phnL* ($\theta_s = 0.136$; $TD_{syn} = 0.33$; RNA-Seq level = 1.4). **f**, Probability density distributions of θ_s values: observed (black) and expected under a uniform mutation rate alone (red). **g**, Fitted lines showing the decay of the correlation between pairs of θ_s values with their genomic distance. Neighbouring genes tend to have similar θ_s, particularly if they share similar functions and they are encoded on the same strand (red line). Weaker but significant correlations are apparent even for functionally unrelated genes (blue) and those on opposing strands (dotted lines). An analogous pattern is observed for θ_s' (Supplementary Fig. 22).

mRNA-folding stability in the 5′ end of genes[16,17], does not explain the variation of θ_s either (Supplementary Information, section 3.1.2).

In addition to direct selection, positive and purifying selection at linked sites (known as hitchhiking and background selection, respectively) affect local neutral diversity in eukaryotes[18]. However, these processes have no effect on synonymous diversity here. Hitchhiking would at most result in lower θ_s in regions under positive selection, which is inconsistent with our observations (Supplementary Information, section 3.1.4). Furthermore, in bacteria where recombination occurs without crossovers and instead resembles gene conversion[19], partial linkage is conserved throughout the entire genome (Supplementary Figs 11 and 16) and background selection cannot preferentially reduce local θ_s values (Fig. 2c and Supplementary Information, section 3.1.4).

The results so far indicate that, contrary to d*S*, the genome-wide variation of θ_s is largely neutral with respect to the strongest known selective forces at synonymous sites in bacteria and that much stronger forces are needed to bias θ_s significantly. Nevertheless, to account for any possible source of selection, including unknown ones, we applied neutrality tests based on the site-frequency spectrum (Tajima's *D*, and Fu and Li's *D* and *F*), which provide relative measures of selection that are independent of the local mutation rate (Fig. 2d)[19]. Using realistic forward simulations of genomes undergoing non-crossover recombination and variable amounts of selection, we studied the impact of selection on θ_s and the allele spectrum. The tests show that Tajima's *D* would detect any traces of selection strong enough to bias θ (Supplementary Information, section 3.1.3).

Regression on Tajima's *D* at synonymous sites (TD_{syn}) indicates that selection from all sources explains less than 5% of the variance in θ_s (Fig. 2e), even accounting for the noise in estimating TD_{syn} using few polymorphisms (Supplementary Fig. 9). This is in agreement with the independently obtained observations above that CUB and mRNA-folding stability have minimal impact on the observed variation of θ_s.

We also explored the influence of non-selective factors on θ_s, including GC content and within-species homologous recombination. All potential biases together only account for 20% of the variation in θ_s (Fig. 2e), suggesting that there is a large underlying variation in the neutral mutation rate along the *E. coli* genome.

Despite the weak association between the above factors and θ_s, we eliminated these small biases using locally weighted scatterplot smoothing (LOWESS) regression for 2,659 genes (Fig. 2f, Supplementary Fig. 19 and Supplementary Information, section 3.4) to avoid ambiguity in the interpretation of the functional associations below. We challenged the regression approach using a cross-validation test (Fig. 3e, top panel, and Supplementary Information, section 3.1.3.6) and extensive forward simulations (Supplementary Information, section 3.1.3.8). They showed that selection bias on θ_s can be comprehensively removed by regression on TD_{syn} even in evolutionary regimes with strong linkage disequilibrium, gene conversion, non-random sampling of bacterial strains and stronger selective forces than those observed here (Supplementary Information, section 3.1.3, and Supplementary Figs 12–14). Thus, the genome-wide variation of the corrected θ_s' values can be considered effectively neutral.

To assess whether the genome-wide variation in mutation rate shows signs of evolutionary optimization, we studied the association between θ_s' and the strength of purifying selection on the protein sequences (TD_{non}). Figure 3a provides initial evidence of the association between the neutral mutation rate and selection: namely, that proteins under stronger purifying selection experience a lower rate.

We then studied the relationship between the mutation rate and functional genomic data for *E. coli* K12. Genes with lower θ_s' tend to be essential for survival in rich media[20] compared with those of higher θ_s' (Fig. 3b). Furthermore, mutationally cold genes generally encode for vital cellular functions such as processes related to central energy metabolism and the respiratory chain (Supplementary Table 1). Mutationally hot genes are associated with metabolic pathways expressed at lower

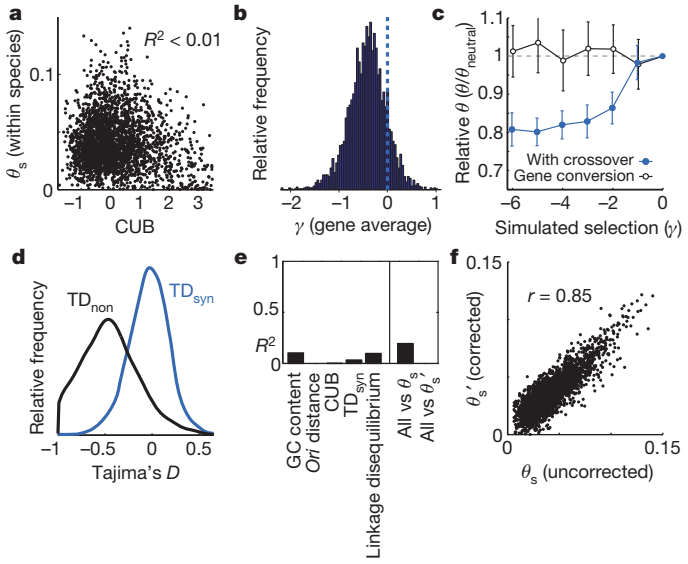

Figure 2 | Selective and non-selective factors have only small effects on the variation of θ_s. **a**, Within-species θ_s does not display the traditional strong correlation with codon usage bias. Only a small fraction of genes with extremely high CUB display a certain decrease in θ_s from selection on codon usage. **b**, Histogram of the weighted average selection coefficient for each gene. These values of γ are too small to cause significant deviations of θ_s (see also Supplementary Fig. 8). **c**, Effect of background selection on the relative local neutral diversity in forward simulations. Background selection cannot affect the local variation of neutral diversity in the presence of non-crossover recombination. Circles represent median values and error bars show the 95% confidence interval of the medians. **d**, Probability density distribution of normalized Tajima's D at synonymous (TD_{syn}; blue) and at non-synonymous sites (TD_{non}; black) across all genes. The distributions indicate strong purifying selection at non-synonymous sites relative to synonymous sites. **e**, Bar plots displaying the proportion (R^2) of the variance of θ_s explained by five selective and non-selective potential biasing factors. All factors combined explain around 20% of the variance in θ_s. Their effect is eliminated from θ_s' after LOWESS correction. **f**, Scatter plot showing the correlation between θ_s and θ_s'. Removal of the potential biases has only a moderate effect on the overall variation of θ_s. Analogous results are obtained with or without the adjustment.

levels or used less frequently, such as amino-acid biosynthesis and catabolism of specific compounds. Interestingly, hot genes do not include antigens and other genes expected to experience frequent positive selection; in fact, we were unable to detect any association between θ_s' and the density of sites under positive selection (Supplementary Information, section 6.4, and Supplementary Fig. 23). Thus, together our observations suggest that purifying, rather than positive, selection has driven the evolution of the mutation rate along the core *E. coli* genome.

Next, we studied the relationship between mutation rate and gene expression using transcriptomic (RNA-Seq) and RNA-polymerase-binding data (Supplementary Information, section 2.4). There is a clear negative correlation between θ_s' and transcription levels (Fig. 3c), with the coldest genes (bottom 5% θ_s') showing nearly fourfold higher transcription than the hottest (top 5% θ_s'). Earlier studies have reported associations between dS and expression levels, but they were generally accompanied by correlations with CUB reflecting the action of selection[13]. Other studies accounting for CUB also suggested that the mutation rate is lower at highly expressed genes[21], but the roles of additional sources of selection remained unresolved. Our data now demonstrate this in the absence of confounding biases.

Finally, because functional properties are often related to expression levels, we performed a multiple regression analysis. This confirmed that the mutation rate independently associates with expression levels, selection on protein sequence and gene function (Supplementary Information, section 6.7).

Given that the process of transcription is known to be mutagenic[22–24], the negative association between expression and mutation rate is unexpected. This implies the presence of compensatory mechanisms that preferentially protect or repair highly expressed loci in *E. coli*. Indeed, we do detect the action of the transcription-coupled repair pathway[25] (Fig. 3d). However, this mechanism alone is insufficient, because the non-transcribed strand also displays a dependence between expression and mutation rate (Fig. 3d), and molecular experiments have reported that transcription-induced mutagenesis occurs in the presence of transcription-coupled repair. Therefore additional mechanisms that generally target highly expressed genes, but are not directly coupled with the transcriptional machinery, must exist.

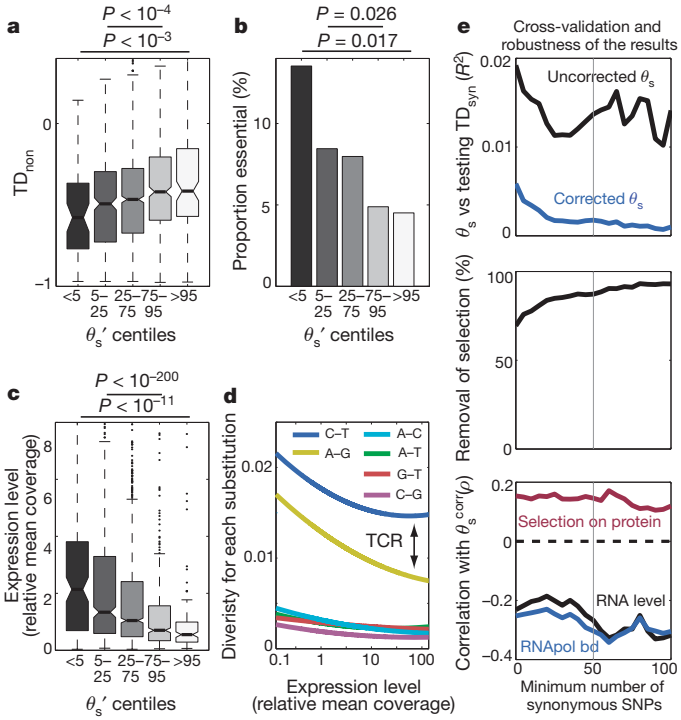

Figure 3 | Variation in the mutation rate shows functional dependence. **a**–**c**, Box- and bar-plots displaying the relationship between gene function and mutation rate. Genes were classified as displaying from very low (less than the fifth centile) to very high (greater than the 95th centile) θ_s' values. P values correspond to (**a**, **c**) Wilcoxon rank-sum test and (**b**) Fisher's exact test. **a**, Genes with higher mutation rate tend to experience weaker purifying selection at the protein level (measured using TD_{non} for genes with at least five non-synonymous single nucleotide polymorphisms). **b**, Genes with lower mutation rate show greater tendency to be essential for survival in rich media. **c**, Genes with lower mutation rate are generally more highly expressed (using RNA-Seq data from *E. coli* K12, Supplementary Information, section 2.4). The central mark of each box-plot represents the median, the edges of the box are the 25th and 75th centiles, and the notches are the 95% confidence interval of the median. Whiskers extend to the most extreme data points within the range. Dots show outliers. **d**, Fitted lines showing the relationship between θ_s' per substitution type and expression level. The direction of the substitution is not shown as we used a reversible evolutionary model. Because the C-to-T transition (G-to-A in the opposite strand) is the most common mutation, the gap between the C–T (blue) and G–A (yellow) lines indicates the strand asymmetry caused by the action of transcription-coupled repair (TCR) on the transcribed strand. **e**, Lines showing the results of cross-validation tests challenging the correction for selection bias (Supplementary Information, section 3.1.3.6). Top, the correlation of θ_s and TD_{syn} before and after adjustment by an independent set of TD_{syn} values shows that selection bias is removed. Middle, proportion of selection bias eliminated (R^2_{after}/R^2_{before}), given different minimum numbers of synonymous single nucleotide polymorphisms in the training set. Bottom, strength of functional associations of θ_s values at different levels of correction, showing that results are unaffected by the adjustment. SNPs, single nucleotide polymorphisms.

Our observations suggest that purifying selection has driven the evolution of the local point mutation rate in *E. coli* to reduce the risk of deleterious mutations. This contrasts with most earlier theoretical work that proposed variants of bet-hedging in which frequent positive selection in changing environments leads to the emergence of hypermutators[1–3]. Instead our observations are in line with an evolutionary risk-management strategy[26] in which sustained stronger purifying selection at specific genes favours individuals with preferential protection or repair at these loci, even at a cost of reduced protection of other genes (see Supplementary Information, section 4.3, for a detailed description of the model). In this way, the rate of deleterious mutations in the genome can be efficiently reduced without excessive investment in protection or repair. In addition, this could increase the rate of non-deleterious mutations, so raising the adaptive potential of the population in case of an environmental change.

We can only speculate about the molecular mechanisms underlying the localized reduction in spontaneous mutations. DNA-binding proteins and DNA repair pathways are obvious candidates, especially as there is increasing evidence for locus-dependent control of the latter[27,28]. In eukaryotes, the best-known sources for mutation rate heterogeneity are sequence dependent; however, analysis of sequence-enrichments around our data set of synonymous polymorphisms provided little indication of context-dependent mutagenesis (Supplementary Fig. 24).

Surprising negative correlations between expression and the numbers of substitutions were recently reported in several human cancers, for which both mutational[29] and selective[30] interpretations have been proposed. Understanding the extent and mechanisms of the modulation of the local mutation rate in different organisms will be important for research into evolution, human diseases, mutagenesis and repair.

METHODS SUMMARY

A total of 3,420 one-to-one orthologues present in at least 26 (75%) of 34 *E. coli* strains were identified by a strict reciprocal best BLAST-hit approach. They were aligned using PRANK-F with a codon substitution model. Multiple phylogenetic filters were applied to discard alignments potentially affected by artefacts. Extensive quality controls were performed on the remaining 2,930 alignments. A θ_s value was calculated for each alignment using a codon model (OmegaMap).

CUB was calculated using the codon adaptation index or the fraction of optimal codons. The scaled selection coefficients ($\gamma = 2Ns$) of selection on codon usage were estimated as in ref. 14. To minimize the impact of selection on mRNA folding, all analyses were repeated after trimming the ends of alignments yielding analogous results. Tajima's *D* was used to estimate selection from any source, and the sensitivity of the approach was tested by multiple independent tests including forward simulations (Supplementary Information, section 3.1.3). The variation of θ_s explained by selective and non-selective biases was removed in 2,659 genes using nonlinear LOWESS regression. This set was used in all functional analyses.

The corrections were challenged using cross-validation tests and extensive forward simulations. The cross-validation showed that TD_{syn} biases can be removed to near completion and that increased removal of selection bias does not affect the functional results (Fig. 3e). SFS_CODE was used to simulate multiple populations of genomes comprising 11 loci under different levels of purifying and diversifying selection, and subject to realistic levels of mutation and non-crossover recombination. For background selection and hitchhiking we simulated coding sequences with selection acting only at non-synonymous sites, and quantified its impact on neutral diversity at synonymous sites under crossover recombination or gene conversion. Homologous recombination was estimated independently by linkage disequilibrium and phylogenetic consistency, yielding analogous results.

Received 27 October 2011; accepted 29 February 2012.
Published online 22 April 2012.

1. Kimura, M. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* **9,** 23–24 (1967).
2. Levins, R. Theory of fitness in a heterogeneous environment. VI. The adaptive significance of mutation. *Genetics* **56,** 163–178 (1967).
3. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4,** 24–33 (1994).
4. Sniegowski, P. D., Gerrish, P. J., Johnson, T. & Shaver, A. The evolution of mutation rates: separating causes from consequences. *BioEssays* **22,** 1057–1066 (2000).
5. Tenaillon, O., Taddei, F., Radman, M. & Matic, I. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res. Microbiol.* **152,** 11–16 (2001).
6. Pal, C., Macia, M. D., Oliver, A., Schachar, I. & Buckling, A. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature* **450,** 1079–1081 (2007).
7. Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. Cryptic variation in the human mutation rate. *PLoS Biol.* **7,** e1000027 (2009).
8. McVean, G. T. & Hurst, L. D. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386,** 388–392 (1997).
9. Chuang, J. H. & Li, H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2,** E29 (2004).
10. Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140,** 783–796 (1995).
11. O'Fallon, B. D. A method to correct for the effects of purifying selection on genealogical inference. *Mol. Biol. Evol.* **27,** 2406–2416 (2010).
12. Bustamante, C. D., Nielsen, R. & Hartl, D. L. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* **63,** 91–103 (2003).
13. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134,** 341–352 (2008).
14. McVean, G. A. & Charlesworth, B. A. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74,** 145–158 (1999).
15. Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. Selection intensity for codon bias. *Genetics* **138,** 227–234 (1994).
16. Eyre-Walker, A. & Bulmer, M. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* **21,** 4599–4603 (1993).
17. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324,** 255–258 (2009).
18. Andolfatto, P. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* **11,** 635–641 (2001).
19. Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5,** e1000344 (2009).
20. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Systems Biol.* **2,** 0008, doi:10.1038/msb4100050 (2006).
21. Eyre-Walker, A. & Bulmer, M. Synonymous substitution rates in enterobacteria. *Genetics* **140,** 1407–1412 (1995).
22. Ochman, H. Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.* **20,** 2091–2096 (2003).
23. Beletskii, A. & Bhagwat, A. S. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **93,** 13919–13924 (1996).
24. Klapacz, J. & Bhagwat, A. S. Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. *J. Bacteriol.* **184,** 6866–6872 (2002).
25. Francino, M. P., Chao, L., Riley, M. A. & Ochman, H. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272,** 107–109 (1996).
26. Wagner, A. Risk management in biological evolution. *J. Theor. Biol.* **225,** 45–57 (2003).
27. Tu, Y., Tornaletti, S. & Pfeifer, G. P. DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *EMBO J.* **15,** 675–683 (1996).
28. Hoege, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature* **419,** 135–141 (2002).
29. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
30. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Ackermann, S. Brenner, G. Dougan, A. Eyre-Walker, N. Goldman, B. Lenhard, J. Marioni, J. Parkhill, O. Tenaillon, C. Tyler-Smith and F. Ulhmann for their suggestions during the preparation of this manuscript. The work was funded by EMBL, the Spanish Ministry of Science and Innovation and the Caja Madrid Foundation.

Author Contributions I.M. and N.M.L. conceived the study; I.M. designed and performed the analyses; A.S.N.S. and N.M.L. provided advice; I.M. and N.M.L. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to I.M. (martinco@ebi.ac.uk) or N.M.L. (luscombe@ebi.ac.uk).